

Animation and 3D color display of multiple-variable data: Application to semiconductor design

by Edward J. Farrell
Steven E. Laux
Phillip L. Corson
Edward M. Buturla

The increasing complexity of digital simulations requires more effective techniques to display and interpret the voluminous outputs. Advanced digital processing workstations and high-resolution color monitors permit a wide range of new techniques for use in examining the global characteristics of each output variable and their interrelationships with other variables. In this investigation, animation, 3D display, and multiple-window imaging have been shown to be effective in interpreting multiple-variable data sets, both scalar and vector. These display methods are used in the solution of two specific semiconductor design problems: the avalanche breakdown of an n-MOSFET and an alpha particle hit on an npn transistor. With these techniques the user can more fully utilize the results of these long and costly computations, making these methods a powerful addition to existing techniques for imaging data.

©Copyright 1985 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

Introduction

Digital simulation is a useful tool for design and performance evaluation. The output data often require additional post-simulation operations: numerical, statistical, or graphical. Numerical methods may be used to compute derived quantities, such as the integral of an output variable over a surface. Statistical methods may be used to interpret results from a simulation involving random variables. Graphical methods range from simple *x-y* plots to multiple 3D displays in stereo. Graphical methods can also be used to obtain quantitative information, such as values at selected points or the position and size of selected regions.

As computational resources become more powerful, simulations become more complex and result in larger volumes of data. With current graphical methods the investigator is limited to looking at a few critical output variables and seldom examines the complete set of output data. For example, a typical 2D simulation of a field-effect transistor over a range of drain biases, with a fixed gate bias, results in 120 2D scalar fields (electron densities, hole densities, and potentials) and 80 2D vector fields (electron and hole current densities). For each new level of gate bias, device geometry, or doping profile, a similar volume of data is generated. Using standard graphical plots of contours or line-mesh 3D surfaces, at best a designer can look at the device terminal currents and a few carefully selected distributions of variable values inside the transistor.

Advanced digital processing workstations and high-resolution color monitors permit a wide range of new techniques for examining the global characteristics of each output variable and its interrelationships with other variables. With effective tools to examine the entire output, there is the possibility of improved understanding and the discovery of new physical results that may not be observed in a limited review of selected output variables. The user can also confirm that the modeling is physically correct, that there are no errors in the numerical computation, and that the iterative computation converges rapidly to reasonable results.

The basic display problem is to develop methods so that the user can perceive the spatial relationships of numerical values within a large, multiple-variable data set. Three approaches are implemented in this investigation: animation, 3D display, and multiple-window imaging. Animation uses the short term visual memory of the viewer to capture in his mind an additional parametric dimension of the data. A 3D display presents the additional dimension as a third axis in the 2D image on the display monitor. The capability in high-resolution monitors to have multiple display windows permits different variables and display methods to be used simultaneously in each window. Animation can be combined with 3D display. For example, a time series of 3D arrays can be formed in successive display windows and then rapidly displayed to show the dynamic evolution of the data, or a series of 3D displays based on different 3D rotations can form an animated image of a turning 3D solid.

The object of this investigation is to demonstrate the effectiveness of animation, 3D color display, and multiple-window imaging. The objective is achieved by using the display methods in the solution of two specific semiconductor design problems based on simulation. This approach serves as an example of how display methods can be combined and used sequentially to interpret large, complex data sets. The methods, programs, and display system are independent of the application area; in addition to transistor design, they have been applied to data from medical scanning, seismic exploration, and simulation of diffusion processes. Other potential application areas are aerodynamics, fluid mechanics, mechanical stress analysis, and thermodynamics.

Several useful methods have been developed to display 3D data or structures. Graphical methods of displaying computer-generated data have been presented by Graedel and McGill [1]. Two-dimensional color plots and 3D binary diagrams are used to display ozone concentration in terms of height, lateral distance, and time. Display of 3D measurements with shading has been used effectively by several investigators for medical applications [2, 3]. Colors for multiple structures and stereoscopy have been used by Farrell to enhance the perception of shapes and relative

locations of structures in 3D data [4, 5]. An effective interactive method for 3D visualization of CAD surface models has been developed by Atherton [6]. Animated 3D imaging is a promising new technique. Curtis and Schwieder have used animation very effectively to interpret the dynamics of a nuclear reactor design based on diagrammatic displays of the system [7]. Our approach differs from these earlier investigations in that multiple-variable data (2D, 3D, scalar, and vector) are interactively interpreted using a high-resolution color monitor and workstation.

Data from semiconductor simulations

The FInite ELement Device ANALysis (FIELDAY) program [8-12] is used to simulate the behavior of semiconductor devices. This program solves numerically the semiconductor transport equations in one, two, or three space dimensions for steady-state, transient, and small-signal ac operating conditions. The program input consists of the device geometry, the voltage and/or current excitations applied to the electrodes of the device, and the impurity atom distribution (doping) within the device. The program generates the internal electron, hole, and potential distributions. From these three output variables, vector fields of hole and electron current densities are computed. These internal current densities are integrated along a device electrode to derive the terminal current entering the device at the electrode. This terminal current can be compared directly to terminal currents measured on laboratory devices.

Typical of a finite element program, the amount of data generated by FIELDAY to quantify a given device design is very large. The independent variables are defined over a 2D or 3D device geometry discretized nominally into 1000-5000 nodes, with even more nodes needed in the future. The corresponding 3000-15 000 output variable values (electron and hole densities, and potentials) only represent one operating condition of the device. To minimally describe the steady-state response of a device to various terminal excitations may require 200 individual operating points. Usually the internal device solutions are checked at certain key operating conditions, but only the terminal current is examined over the entire range of simulated data. The use of contour or other plot types to attempt a more global interpretation of internal device variables is unwieldy and cumbersome. However, such global interpretation of these data may well lead to enhanced and unified insight into device operation.

Here animation, 3D display, and multiple-window imaging are used in the solutions of two semiconductor design studies based on FIELDAY simulations. The first study is a parametric 2D investigation of channel breakdown in an n-MOSFET. Three gate potentials are simulated, each with a wide range of drain potentials. The goal of the n-MOSFET simulation is to better characterize the breakdown

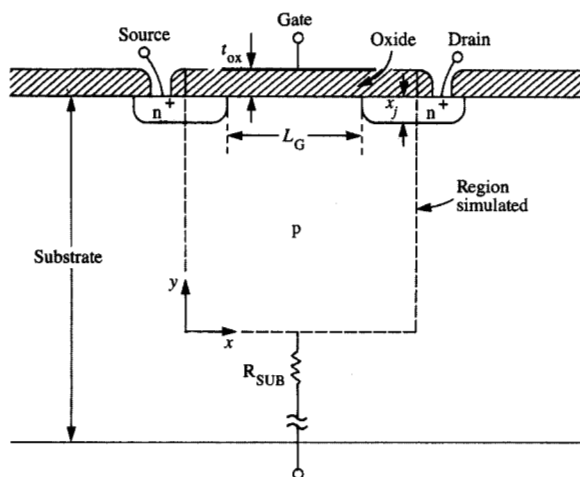


Figure 1

Cross section of an n-MOSFET showing the region simulated.

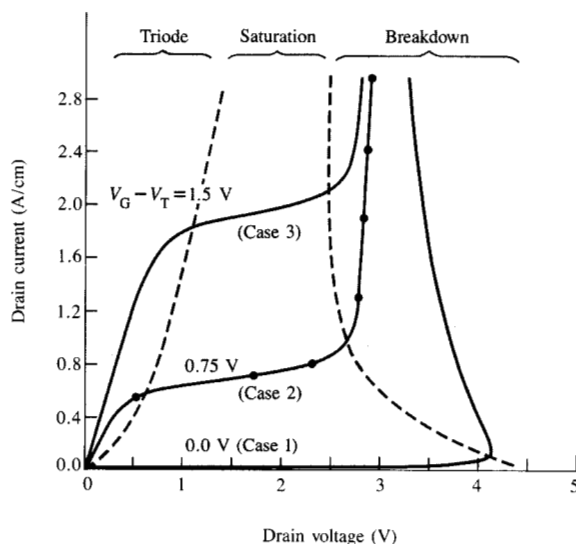


Figure 2

n-MOSFET drain current vs. drain voltage for three levels of gate potential. Three operating regions are shown: triode, saturation, and breakdown. The six points are referenced in the section "Multiple-variable data in 2-space."

process. The second study is a series of 3D FIELDAY simulations over time which model the effect of an alpha particle hit on an npn bipolar transistor. The goal of the 3D simulation is to understand the mechanism of charge redistribution and collection within the device after the alpha particle hit.

• n-MOSFET simulation in 2-space

The most pervasive semiconductor technology for integrated circuit fabrication is that of the n-channel silicon metal-oxide-semiconductor field-effect transistor (n-MOSFET). Pioneering work on the scaling of n-MOSFETs [13, 14] provides an algorithm for n-MOSFET miniaturization accompanied by improved switching performance. Unfortunately, the power supply voltage is not always scaled according to this algorithm. As a result, n-MOSFETs with channel lengths in the micron and submicron range are internally subjected to more high energy carriers than past supermicron n-MOSFET technologies. One important physical process present in such cases is carrier generation by impact ionization, which leads to avalanche breakdown. An investigation of carrier generation and eventual breakdown in an n-MOSFET serves as a vehicle to show the utility of animation, 3D display, and multiple-window imaging.

A brief description of n-MOSFET operating principles is appropriate to acquaint nonexperts with the basic operation of the device. The n-MOSFET to be examined in this paper is shown in cross section in **Figure 1**. The primary conduction mechanism in an n-MOSFET is the flow of n-type carriers (electrons) from the electron-rich n⁺-doped pocket at the source to the electron-rich doped pocket at the drain. The gate potential modulates the strength of the conduction path; the potential of the substrate contact also weakly modulates this path. When the gate voltage is below a critical level (the threshold voltage V_T), the conduction path is ineffectual because the hole-rich p-type substrate blocks the conduction path.

In our FIELDAY simulation study, a complete range of drain currents and voltages were computed for three values of gate voltage: *Case 1*, gate voltage equal to V_T ; *Case 2*, gate voltage 0.75 V above V_T ; *Case 3*, gate voltage 1.5 V above V_T . The device region simulated is outlined by a dashed line in Fig. 1. For Case 1, 39 drain potentials were simulated; for Case 2, 47; and for Case 3, 33.

The output characteristic generated by FIELDAY is shown in **Figure 2** and is divided into three operating regions: triode, saturation, and breakdown. For Case 1, very low values of drain current are present out to the breakdown region. For Cases 2 and 3, the positive gate potential pushes holes away from the surface and allows an electron conduction path to form. As the drain voltage increases, drain current increases rapidly and linearly due to a low channel resistance resulting from a heavy accumulation of electrons at the substrate-gate oxide interface. The drain current does not increase without bound, however. The conductivity of this path depends on the potential drop across the gate oxide; the greater this voltage drop, the lower the conductivity of the channel. As the drain voltage increases, the potential drop across the gate oxide decreases at the drain end of the channel, and the electron channel weakens and disappears near the drain. This condition is

known as channel pinch-off. The pinch-off region is a low-conductivity interval near the drain in series with the highly conductive channel. As a result of pinch-off, saturation occurs and the I - V curves bend over as shown in Fig. 2. These aspects of triode and saturation region operation are well documented [15, 16].

Breakdown in the n-MOSFET is initiated when excess holes are created near the drain by impact ionization; these holes flow out of the contact at the bottom of the substrate in Fig. 1. Hole flux causes a significant potential increase near the source, which triggers parasitic electron injection from the source, and breakdown occurs. The advanced nature of the n-MOSFET simulation used in this study is reflected by its inclusion of the breakdown process. Detailed discussions of device breakdown have appeared [17-19], but only one fully self-consistent simulation in the region of negative slope present in breakdown (snapback region) has appeared [20].

The n-MOSFET device is simulated in the x - y plane and is assumed invariant in structure in the z -axis direction. The substrate effects deep below the device are modeled as a lumped resistance R_{SUB} . The modeled device has a gate length $L_G = 0.5 \mu\text{m}$ and gate oxide thickness $t_{ox} = 12.5 \text{ nm}$. The substrate is doped p-type with acceptor doping density $N_A = 6 \times 10^{15} \text{ cm}^{-3}$. A channel implant is assumed which yields a threshold voltage $V_T = 0.4 \text{ V}$. Source and substrate biases are $V_S = 0 \text{ V}$ and $V_{SUB} = -0.1475 \text{ V}$, respectively.

The 2D grid used for simulating this n-MOSFET consisted of a highly nonuniform finite element mesh with 962 nodes and 1822 elements. The problem occupies 5.45M bytes of IBM 3081 memory; on average, one drain bias point requires six minutes of CPU time. Thus, to generate the three curves shown in Fig. 2 requires 720 CPU minutes. A total of 2.8M bytes of potential, electron, and hole data are computed. The volume of current density and generation/recombination data for these studies is also several megabytes. For graphical processing, these data were mapped onto a uniform rectangular grid of 118×128 pixels. Potential was linearly interpolated onto this grid; electron and hole densities were logarithmically interpolated. For complex vector fields, plotting arrows throughout the mesh does not simultaneously portray magnitude and direction accurately; thus a new approach was taken. Two separate data arrays were computed, one depicting vector magnitude and one depicting vector direction. Vector magnitude is a scalar; hence, it is readily calculated and interpolated onto the graphical mesh. Vector direction is represented by lines which follow the vector field direction. Since electron and hole sources and/or sinks are conditionally present in the model, the lines may start or stop within the device. The direction lines are initiated from points selected along a user-specified line segment. These points are determined to equidistribute the line integral of the vector component transverse to the line segment. This method of selecting

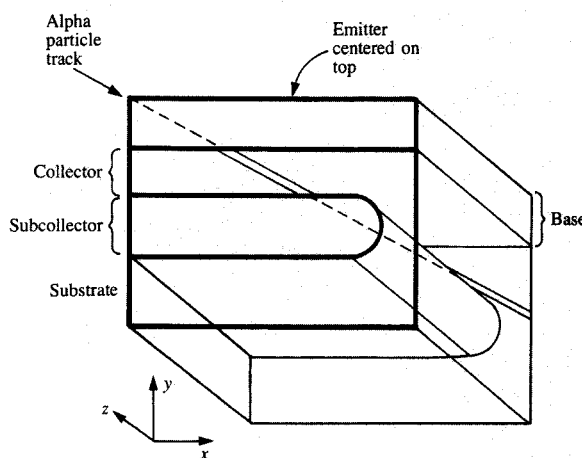


Figure 3

Diagram of simulated bipolar npn transistor. Alpha particle track enters on the upper left at nine degrees to the horizontal plane.

starting points has performed satisfactorily; further, if all sources and sinks are removed from the simulation, it results in true streamlines obtainable from a harmonic function.

• Bipolar transistor simulation in 3-space

As physical dimensions of devices shrink to improve performance and/or density, logic errors attributable to ionizing alpha particles become more severe. Such errors occur because the number of mobile electrons and holes introduced into a structure by an alpha particle hit can be comparable to the total number of carriers ordinarily in the device. Such a flood of unwanted additional carriers within the device can literally cause a memory state to be changed. The error will not be permanent, however, as no lasting damage is incurred by the device. These errors are accordingly termed "soft errors" [21]. An alpha particle hit on an npn bipolar transistor is modeled in 3D using FIELDAY. The goal of the simulation is to understand the mechanism of charge redistribution and collection within the device after the hit.

The 3D device geometry used for the simulation is shown in Figure 3. The emitter and base contacts are on the top. The subcollector contact is at the side, while the substrate contact is at the bottom. The alpha particle path is nine degrees to the surface as shown. The simulated region is $6.0 \mu\text{m}$ (x direction) by $4.82 \mu\text{m}$ (y direction) by $6.0 \mu\text{m}$ (z direction). The base width is $0.25 \mu\text{m}$. Throughout the simulation the device is biased as follows: collector bias $V_C = 0.75 \text{ V}$, base bias $V_B = 0.40 \text{ V}$, and emitter bias $V_E = 0 \text{ V}$. The p^- substrate below the subcollector is biased at -1.5 V .

The grid used for simulation of the device consisted of a regularly spaced triangular mesh with 4864 nodes and 8100

elements. The x , y , z dimensions of the output array were 16, 19, 16, respectively. This was the densest grid that could be used because of CPU time and memory constraints. The steady-state response of the device prior to the alpha hit was first calculated. The alpha particle hit is assumed to be instantaneous; the internal state of the device was then examined for fifteen logarithmically spaced time steps. The shortest time step was 1.0 ps; the longest was 5 ns. The total time simulated was 10 ns. This simulation scenario required about 1000 CPU minutes and 45 megabytes of memory, running on an IBM 3081 processor under the MVS XA operating system. The output of the simulation program was a data set containing the three 3D scalar arrays (electron and hole densities, and potential) at each time point of the simulation.

Current methods for displaying these 3D simulation results are line-mesh 3D surfaces or contour plots of each 2D x - y plane. This involves 720 plots (three variables $\times$ 15 z planes $\times$ 16 time steps) which must be mentally combined by the designer to draw a conclusion regarding device behavior. Because of this difficulty in interpreting the results, device simulation in three dimensions has been of little practical value to the designer.

To image the output arrays with animation and 3D display methods, the number of points in x , y planes were expanded using a bilinear interpolation function; x , y dimensions of the expanded arrays were 151 and 109, respectively. The logarithms of hole and electron densities were used for interpolation; potential was used without a scale transformation.

Multiple-variable data in 2-space

Several types of simulations and measurements result in a series of 2D arrays of scalar or vector variables. The series may be the result of a simulation of a 2D process for a range of parameter values or the result of 2D scan measurements taken at different physical locations. To illustrate the effectiveness of animation, 3D display, and multiple-window display for the interpretation of a series of 2D arrays, a specific semiconductor design problem is addressed. The task is to characterize avalanche breakdown of an n-MOSFET based on FIELDAY simulation studies. As described in the section "n-MOSFET simulation in 2-space," the simulation output consists of three scalar arrays of electron density, hole density, and potential. Two vector arrays of current flux and hole flux and a scalar array of net recombination rate are computed from the three primary output variables.

• Animation

Since the data are two-dimensional, it is useful to examine them initially with 2D animation. This permits the user to scan the data for consistency and to perceive how the data change with the parameter value. Since all of the data arrays in a series are stored in the imaging workstation, they can be

rapidly displayed to form animation without special operations. Six frames from four animation series are shown in Figure 4 for Case 2 (gate potential 0.75 V above V_T). The six corresponding points on the middle I - V curve are shown in Fig. 2.

The first series (a) is drain potential, with the red regions corresponding to potentials greater than -0.08 V. The drain (upper right corner) has increasing potential throughout the sequence; the first three frames represent triode and saturation operations. In the fourth frame, the increase in potential at the source (upper left corner) is evidenced by the growth of the colored region in this vicinity. The final two frames show strong breakdown with the potential roughly constant over most of the substrate region.

Series (b) and (c) show hole density, with yellow representing densities greater than $6.8 \times 10^{14} \text{ cm}^{-3}$. The contouring in series (c) is similar to contour lines; they are formed by simply quantizing the color changes with a step function. In most cases contouring aids in the interpretation of data during animation. The first two frames show that as drain bias rises, the drain depletion region increases and pushes holes away from the drain. In the third and subsequent frames, holes created near the drain end of the channel are observed. These holes are created by channel electrons which gain sufficient energy near the drain to cause impact ionization. The holes move towards the bottom of the n-MOSFET, resulting in a significant hole current in the substrate. As the drain potential increases, the avalanche region around the drain increases and the region of hole flux broadens. Due to resistive drops in the device substrate (both distributed and lumped), this hole flux results in an increase in potential near the source. This allows parasitic electron injection from the source, creating a subsurface current, as shown in the fourth and subsequent frames of the series (d) in Fig. 4. In this series, eleven electron current density flow lines are superimposed on electron current density magnitude (green indicates magnitude above $3.4 \times 10^3 \text{ A/cm}^2$). This pronounced increase in drain current marks the beginning of the breakdown region.

The flux image frames (direction and magnitude) are formed as the data are loaded into the workstation storage buffers. The input magnitude data are scaled to lie between 0 and 253 and are temporarily written into the refresh buffer. The flux lines are then written into the same window of the refresh buffer using a hardware vector generator. The lines are written as 254. This amplitude encoding permits the user to display flux lines only, flux lines and magnitude, or flux magnitude only with a few missing points at line locations. The 3D axes are written as 255 and may be displayed or dropped in the 3D image. The refresh buffer image (magnitude, direction lines, and axes) is loaded into the storage buffer of the workstation. The data have been scaled between 0 and 255 because the storage buffers are 8 bits per pixel.

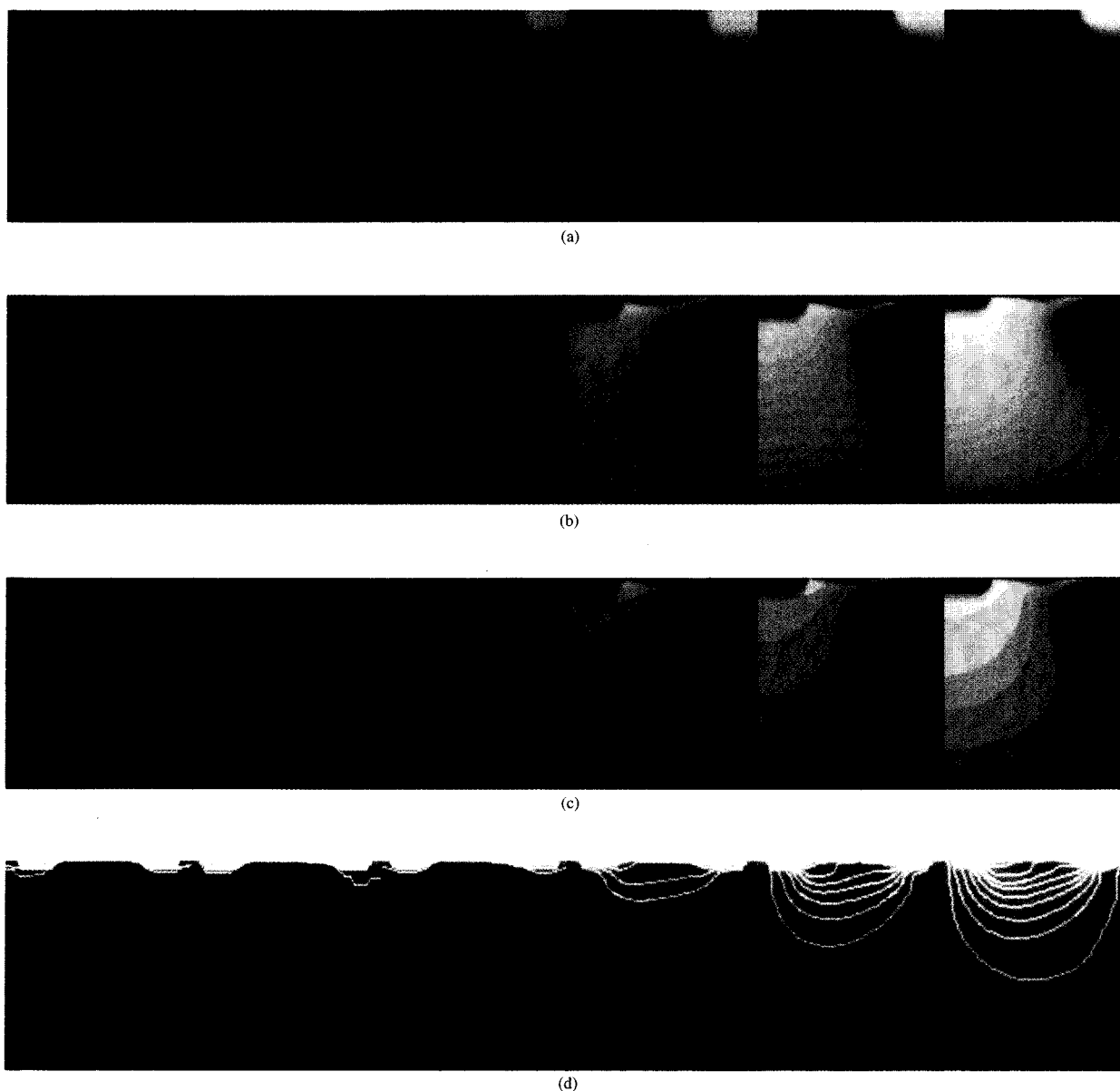


Figure 4

Animation sequences for the n-MOSFET Case 2 simulation: (a) drain potential, (b, c) hole density without and with contouring, and (d) electron flux showing magnitude and direction lines.

• 3D display

A 3D display is an alternative method to present data like those shown in Fig. 4. Successive 2D frames are overlaid on the CRT so that the axis coming toward the viewer represents the parametric change. In the n-MOSFET application, the user can see in one image the changes in a selected internal variable over the entire span of drain voltages at a fixed gate bias. A 3D display is particularly helpful when several variables are compared or when a single image of the data is required.

The 3D display is based on a colored-range method which represents all points whose values lie in a specified range as a single color. For example, all points whose values lie between 10 and 57 are represented as red points, and all points whose values lie between 79 and 105 are represented as green points. A 3D display is then formed by sequentially projecting the frames with the colored ranges onto the CRT display. Different rotations are formed by using different projection parameters. The distance of the frames from the observer is indicated by displaying closer points with lighter



Figure 5

Electron density for Case 1, stereo pair. The orange region corresponds to densities above $8.4 \times 10^{15} \text{ cm}^{-3}$.

colors of the same hue, e.g., lighter red and lighter green. The colored-range method was described earlier in detail by Farrell [4]. The back-to-front increase in lightness results in pseudo-shading. The details of the data processing and management are discussed in Appendix A.

The colored-range approach has four advantages. First, several value ranges can be displayed in the same 3D image using different colors. Second, as successive frames are overlaid on the CRT display, the relative position and size of various value ranges can be interpreted; information is obtained during the development of the final 3D image. Third, it is fast, since the shading is not based on computing surface normals with light source and viewer directions. Also, the speed is independent of which options are used (except for stereo, since two images are formed). A fourth key advantage of this approach is that data need not be continuous. Very irregular value regions can be displayed, as well as a disjoint collection of points.

Because of the complex shape and interrelations of various colored regions (value ranges) and direction lines, a single presentation mode is not sufficient. Five display options can be used to enhance the perception of the three-dimensional relationships: rotation, cutout, transparency, contours, and stereoscopy. Also, the interactive selection of the colors used for the display can significantly improve its clarity, as described in Appendix B.

By rotating the 3D array of colored ranges and direction lines, the user can obtain an unobstructed view of important spatial relationships. The cutout, transparency, and contour options yield new information that cannot be obtained from a single image regardless of rotation. A cutout is formed using write-masks; on each frame near the observer a mask is used to select a specified region. The mask region is entered with a joystick over a selected image for reference. Transparency is formed by writing alternate image points in

a checkerboard manner. This method results in a true admixture of colors. Surface contour lines are formed by using a contrasting color in selected rows and columns of each 2D frame. After forming the projection with these frames, the contour colors appear as surface contour lines in the 3D display.

With stereoscopy a cluster of colored regions and flux lines can be easily related spatially in 3D. A stereoscopic pair is generated by forming two images with different rotations about the vertical axis, usually 10 to 20 degrees apart. A reflecting stereoscope is used to view the images; it diverts the lines of sight so that each eye only sees the correct image for perception of stereoscopy. Cross-eyed viewing will also work.

A stereo 3D display of electron density for Case 1 is shown in **Figure 5** and relates channel pinch-off, subsurface conduction, and breakdown. The drain potential increases toward the rear; the drain is in the upper left corner. Pinch-off is displayed as an ever-widening gap in the colored region in the upper left sector. Note that pinch-off occurs for all drain potentials. Roughly two thirds of the way back along the receding axis this gap appears to close; however, this is not the case. Rather, pinch-off has strengthened to the point where the electron density curves down around the pinch-off point and colors in the gap region. In the rearmost slices of electron density, the parasitic injection of electrons from the source into the substrate is readily seen. This injection increases in strength and forms a subsurface conduction path between source and drain. In the final slice, virtually all of the device is flooded with electrons injected from the source.

The hole density and electron flux shown in Fig. 4 can be combined into a single 3D display to clarify their relationship. **Figure 6** is a stereo 3D display of electron flux lines and hole density for Case 2 with the same orientation as Fig. 5. This figure shows that in the breakdown region the

subsurface electron conduction (the flux lines) intercepts the region of large hole density around and below the source. A direct result of this intersection of electron and hole fluxes can be seen in a stereo 3D display of total recombination rate, as shown in Figure 7. In this figure the slices closest to the front are not colored; only low-level thermal generation/recombination is present in the triode region and the beginning of the saturation region. Impact ionization intensifies with increasing drain potential, as shown by the growth of the red region. Recombination becomes a more pervasive process as breakdown is encountered. Recombination is most intense where electron and hole concentrations are simultaneously high. This occurs precisely around and below the source, as predicted by Fig. 6. In the

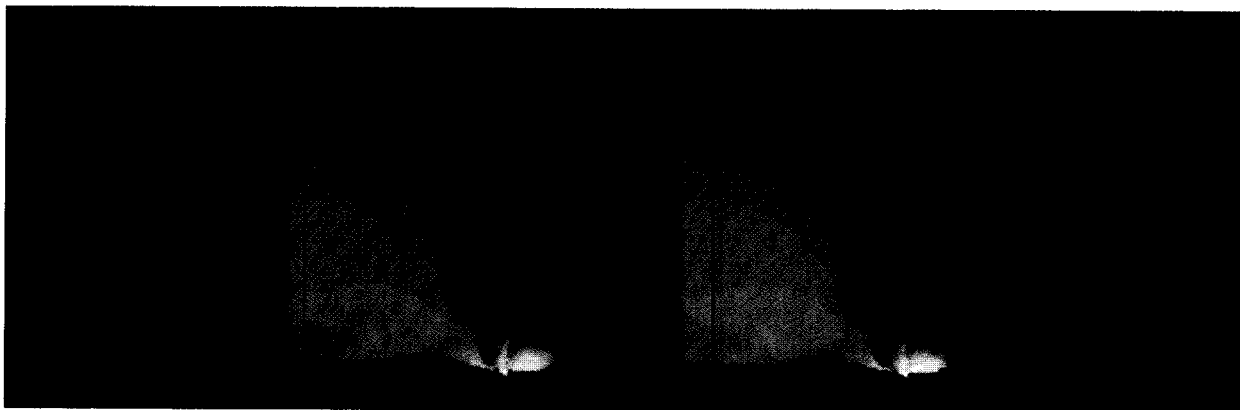
final slices, as the device floods with excess electrons and holes, recombination is present throughout the substrate.

- *Multiple-window imaging*

Animation and 3D display are useful methods for displaying a single variable. To compare several variables these methods must be combined with multiple-window imaging. This technique is based on forming images in different windows in the workstation refresh buffer. With a hardware zoom capability any window can be instantaneously displayed on the full monitor screen by changing the logical origin of the refresh buffer. Since the refresh buffer is square, with 1024 points on a side, 64 windows 128×128 can be defined with a zoom of 8, or 16 windows 256×256 with a

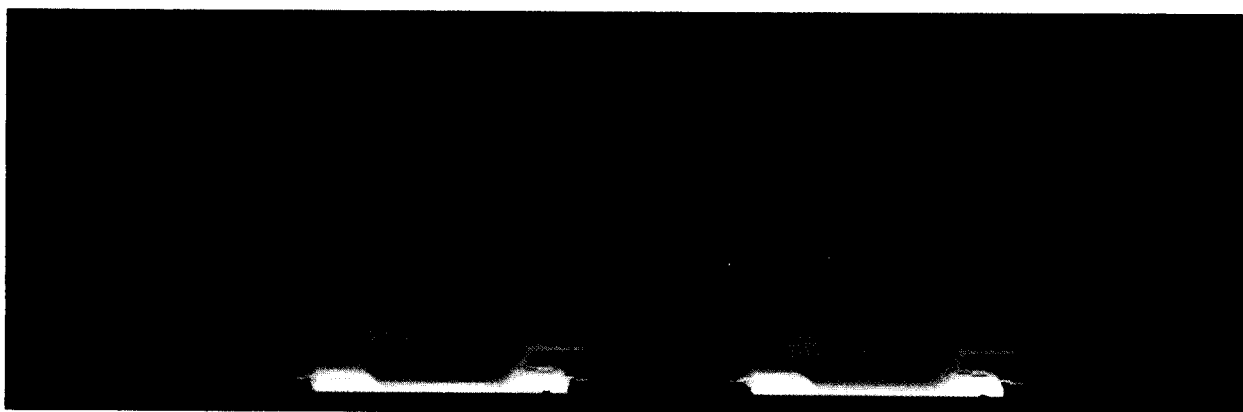
Total recombination rate for Case 2, stereo pair. The red region corresponds to generation rates in the range of 8.7×10^{16} to $1.0 \times 10^{30} \text{ cm}^{-3} \text{ s}^{-1}$; the orange region to recombination rates from 1.0×10^{20} to $1.0 \times 10^{26} \text{ cm}^{-3} \text{ s}^{-1}$.

Figure 7



Electron flux lines and hole density for Case 2, stereo pair. The orange region corresponds to densities above $1.3 \times 10^{16} \text{ cm}^{-3}$.

Figure 6



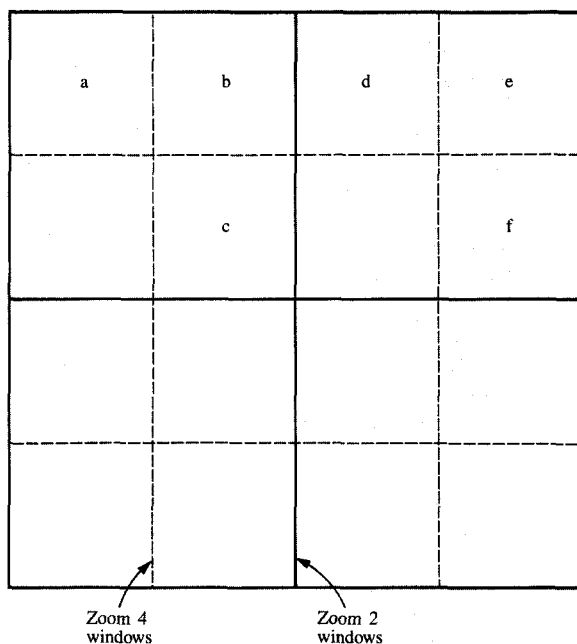


Figure 8

Refresh buffer window definition.

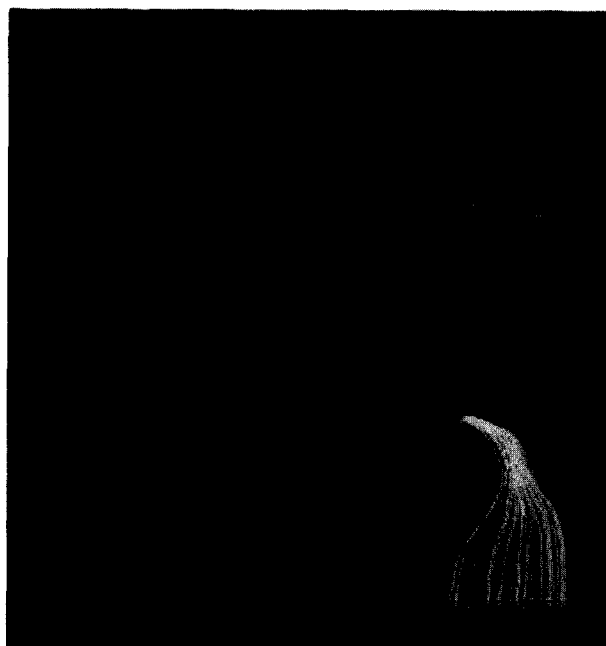


Figure 9

Hole flux lines for Cases 1, 2, and 3, counter clockwise from lower right.

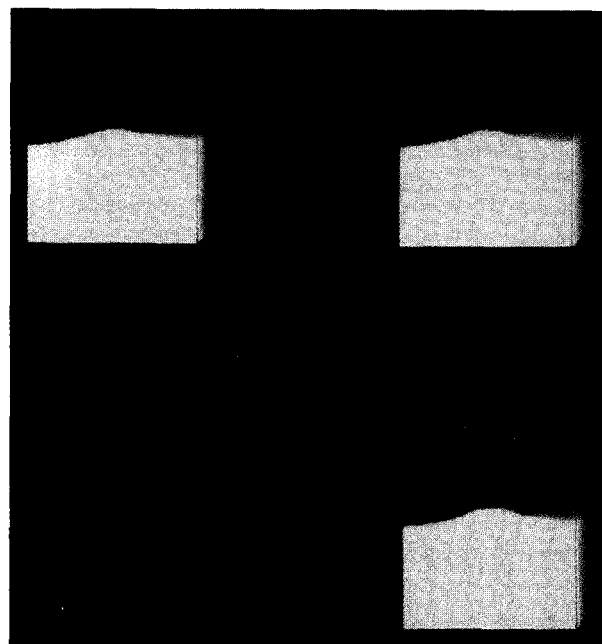


Figure 10

Hole density for Cases 1, 2, and 3, counter clockwise from lower right. The colored region is for hole density greater than $4.2 \times 10^{15} \text{ cm}^{-3}$.

zoom of 4, or 4 windows 512×512 with a zoom of 2. The windows may contain 2D or 3D images, series of images to be displayed in animation, graphic plots, or text.

Multiple windows are required to further characterize the avalanche breakdown of an n-MOSFET. It is helpful to compare the response of the device to increases in drain potential for several gate biases. Two variables are imaged: hole flux lines and hole density. While the images are being formed, the refresh buffer is divided into 16 windows as shown in **Figure 8**. Hole flux lines for Cases 1, 2, and 3 are written into locations a, b, and c, respectively, and hole density into locations d, e, and f. For comparison of these different sets of 2D displays, the zoom factor is set to 2; the refresh buffer is divided into four windows. Each window contains all the data generated for the three cases. **Figure 9** is the upper left quadrant, and **Figure 10** is the upper right quadrant. The user can instantaneously switch between sets of 3D displays.

Figure 9 depicts two interesting results. First, the first slice colored for Case 1 is lighter than for Case 2 and both are lighter than for Case 3. This signifies that hole generation begins closest to the viewer for Case 1 and farthest from the viewer for Case 3. This is simply a reflection of the fact that more $I-V$ data are generated for Case 1 after the onset of hole generation, followed by Cases 2 and 3, respectively. Second, in the rearmost slices, remarkably similar flow paths are seen near the source (upper right), but very dissimilar

path shapes are seen near the drain (upper left). The rise in potential below the source allows a very similar path to be taken by holes near the source. However, the downward-pointing hole flux path below the drain for Case 1 is weakly evident in Case 2 and not present in Case 3. The presence of this path depends on the impact ionization of electrons flowing in a subsurface channel and entering the drain from below. This only occurs noticeably in Case 1, as the fraction of total drain current flow due to breakdown is much larger than for the other two cases; thus, the subsurface flow is the most intense of the three cases.

Figure 10 shows hole density for Cases 1–3. This figure contrasts the substrate hole densities (lower half of 3D images) depleting away from the drain (upper left) with increasing drain bias. Also, the images illustrate that holes are created near the drain and flow towards the back side of the device. The three images compare nicely the intensity of this hole generation. The final slice of the image for Case 3 has the weakest hole generation of the three. This results in a thin strip of holes extending downward from the drain end of the channel. Case 1 has the strongest hole generation in its last slice. Generated holes move in a broad path away from the drain end of the channel. In contrast to Cases 2 and 3, the path has widened sufficiently to cause large hole densities at the silicon-gate oxide surface in midchannel. Figures 9 and 10 also indicate that Case 1 has the strongest lateral and downward movement of holes.

Numerical convergence of 2-space simulations

Animation, 3D display, and multiple-window imaging are used to characterize the breakdown process in the section "Multiple-variable data in 2-space." Since these methods can be applied to any multiple-variable data set, they can also be used to study numerical convergence of the simulation itself. In FIELDAY, a modified Newton's method iteration is used

to solve simultaneously the nonlinear semiconductor device transport equations. The path taken by the iteration in moving from an initial guess to converged solution is of interest. This iteration should consume as little CPU time as possible and should converge independently of the initial guess. The 2D arrays of iteration variables can be compared during convergence. Alternate damping algorithms and error cross-coupling between variables can be examined from a 2D viewpoint.

The use of latency in dynamically reducing the size of the matrix being factored has been proposed as one method to reduce CPU time [22]. To best profit from this method, the Newton's correction vector should be highly localized in the 2D device cross section. That is, the iterative method should show markedly slower convergence in localized areas of the mesh. The convergence path is examined for the Case 2 solution with $V_D = 2.5$ V, using the Case 2 solution with $V_D = 2.4$ V as the initial guess. Eighteen iteration steps are required for convergence.

Potential convergence is determined by $q\Delta\psi/kT$, that is, the potential change $\Delta\psi$ normalized by the thermal voltage $kT/q = 0.02586$ V. The potential is considered converged if the absolute value of this function is less than $\epsilon = 2 \times 10^{-4}$ at all nodes. Electron convergence is determined by $\Delta n/n$, that is, the fractional change in the electron density. The electron density is considered converged if the absolute value of this fractional change is less than ϵ at all nodes. Hole convergence is determined similarly to electron density. Eight snapshots from the convergence path for carriers are shown in Figure 11, representing the Newton's corrections for iterations 1, 3, 5, ..., 15. Red regions indicate where the fractional changes in carrier concentrations were positive and above the convergence threshold; green regions indicate areas that were negative and below the convergence threshold. Blue regions are converged.

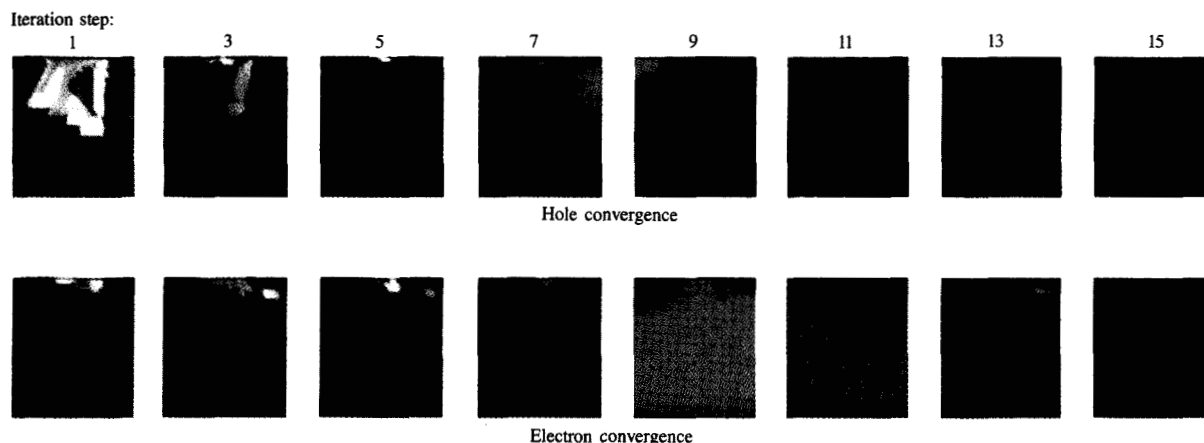


Figure 11

Numerical convergence sequences for electron and hole convergence.

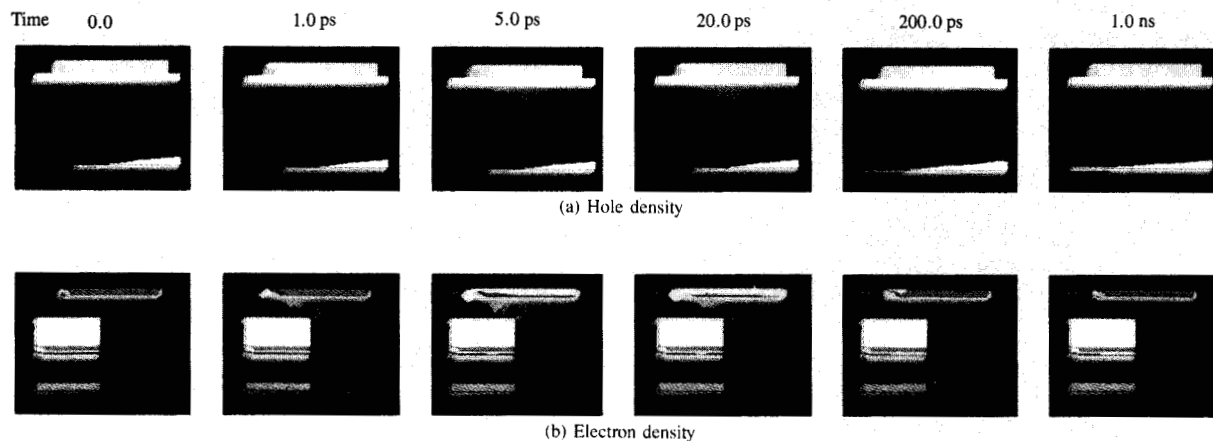


Figure 12

Part of the hole and electron density animation series from the 3D FIELDAY simulation. (a) Orange represents hole densities between 1.4×10^{13} and $8.7 \times 10^{17} \text{ cm}^{-3}$. (b) Electron densities are represented by colored intervals as follows: red, $4.4 \times 10^{19} \leq n \leq 1.1 \times 10^{20} \text{ cm}^{-3}$; orange, $3.1 \times 10^{16} \leq n \leq 1.3 \times 10^{18} \text{ cm}^{-3}$; green, $n \leq 7.3 \times 10^3 \text{ cm}^{-3}$.

When this fractional change criterion is used for convergence, it is evident that convergence proceeds in a nonlocalized, broad-based fashion for both carriers through iteration 7. Beyond this point, holes do converge somewhat more rapidly than electrons, as can be seen. However, not until iteration 13 are a significant portion of the hole values converged. Furthermore, the highly nonuniform mesh has few nodes in the bottom third of the device. For this problem, and using this convergence criterion, the ability to save substantial CPU time by dynamically reducing the matrix size seems improbable.

Multiple-variable data in 3-space

For multiple-variable data in 2-space, animation or 3D display provides an additional display dimension for parametric changes. For multiple-variable data in 3-space, three dimensions are required just to display all of the data. Hence 3D display with animation and multiple-window imaging is essential to view and interpret multiple-variable 3D data. These methods are used to display results from a 3D FIELDAY simulation of an npn bipolar transistor. Three time series of three-dimensional arrays are generated: electron densities, hole densities, and potentials. At each time step a 3D display of an output array presents the spatial relationships of selected value ranges. With multiple-window imaging, one entire series of 3D displays can be stored in the refresh buffer. The series can be reviewed stepwise or displayed in animation to simulate the time evolution.

Part of the series for hole and electron density is shown in Figure 12. The spatial orientation in Fig. 12 is the same as in Fig. 3; that is, the emitter is on top, and the substrate is on

the bottom. The alpha particle enters the simulated region in the upper left corner nearest the front. The first frames in the two series are the solution prior to the alpha hit. The holes located in the base are seen near the top of frame 1, Fig. 12(a). In Fig. 12(b) the electron-rich emitter is seen as red and orange in frame 1; the large orange shell in the middle is the transition from collector to subcollector and subcollector to substrate. A cutout has been made in the lower right corner nearest the front. It permits viewing "inside" the orange shell to see the location of the heavily doped red subcollector. One ps after the alpha particle hit, the alpha path is clearly evident in both hole and electron frames. After 5 ps, the electron-hole pairs created by the alpha hit are less focused around the path as the charge is being redistributed throughout the device. After 20 ps, the electrons are collapsing along the alpha path but are moving deeper into the device; holes are seen to be most prevalent towards the back and right side of the device. After 200 ps, the electron disturbance is limited to the lower far corner where electrons are in the substrate. At this same time, a large hole density is still present in the base-collector depletion region; however, the primary flow path for holes is towards the p-type subcollector. At 1 ns, the electron and hole disturbances are virtually gone.

From this series little evidence for funneling is seen. There are two reasons for this. First, the alpha path passed very close to the subcollector, causing carriers to be collected from below instead of flowing backwards along the alpha path. Secondly, the rough edges in the colored areas of these frames indicate that the mesh used to simulate this problem is still too coarse. For these reasons, the simulation should be repeated with a different alpha path and a finer mesh.

Discussion

The increasing complexity of digital simulations requires more effective techniques to display and interpret the voluminous outputs. The MOSFET simulation presented in this paper produced 2.8M bytes of output data after 720 CPU minutes. Even though the second simulation of an npn transistor produced less than 1M byte of data, 1000 CPU minutes on a 45M-byte computing system were required to generate the data. The object of this paper is to demonstrate the effectiveness and application of animation, 3D color display, and multiple-window imaging to more fully utilize the results of these long and costly computations. The basic design problems are to characterize the avalanche breakdown of the MOSFET and to determine the 3D redistribution of carriers after an alpha particle hit on the npn transistor. We obtained results in five specific areas. First, both the 2D and the 3D FIELDAY simulation results are qualitatively correct, with no computational anomalies for the range of device biases simulated. Second, for the 2D MOSFET simulation, the concise display of the simulation results for a wide range of operating conditions provides enhanced understanding of MOSFET breakdown. Third, a multiple-window image of the MOSFET simulation convergence shows that no significant reduction in CPU time can be achieved through dynamic spatial adaptation of the computational mesh. Fourth, for the 3D FIELDAY simulation of an alpha particle hit, animated images of the 3D distributions of hole and electron densities show that minimal funneling occurs. Fifth, the 3D display of the 3D npn transistor simulation output indicates that results are degraded by the mesh coarseness.

At present, the device designer interacts with the data either in a very quantitative way using a single contour or 3D line-mesh plot or in a very global way by studying the I - V behavior of the device. Animation, 3D display, and multiple-window imaging offer an intermediate option. A range of parameter values is presented with animation or 3D images, yielding a global interpretation of the data without sacrificing a quantitative image of behavior in the 2D device plane. Specifically, with animation (Fig. 4) we are able to confirm that the n-MOSFET model functioned properly and to understand better the mechanism of breakdown in this device. Individual frames of 2D data could be examined in detail as needed, but the true strength of animation is in cycling through a large amount of data quickly. In this way, behaviors spanning multiple 2D frames become readily discernible. As an alternative to animation, or where animation is not readily possible (such as on the printed page), 3D display (Figs. 5 and 6) offers a way to concisely display a span of parameterized 2D data. Again, trends present through multiple 2D frames of the data become clear. Multiple-window imaging (Figs. 9 and 10) aids in understanding the relationships between images *each* representing a large volume of data. Specifically, the variations in hole flux lines and hole density for the entire

I - V curve of Fig. 2 were compared with the multiple-window images in Figs. 9 and 10. These display methods can also be extended directly to evaluate the longitudinal and spatial variations in numerical convergence, as presented in Fig. 11. They are also useful in tuning numerical algorithms to yield maximum performance.

The techniques discussed here for viewing multiple-variable data in 3-space virtually supplant present imaging techniques of such semiconductor simulation data. Representation of 3D data as 3D images instead of as a decomposition of 2D planes is invaluable for the interpretation and understanding of such data. Figure 12 showed the ease with which a complex physical process modeled in 3D could be studied. Although the simulation itself was less successful than originally hoped on physical grounds, the imaging techniques described are shown to be of paramount importance in the interpretation of such complex 3D simulation data.

Animation, 3D display, and multiple-window imaging are applied to two specific semiconductor design problems, but the methods can be applied to a wide range of applications. In complex computer simulation models the entire output can be examined for all parameter values, and not just a few key output variables. With detailed results displayed for evaluation, the simulated device or process can be better understood, with the possibility of discovering new results. Also, the functional performance of the model and its numerical convergence can be evaluated. The results presented in this paper demonstrate the utility of the three imaging methods, and the potential of interactive color imaging for the display and interpretation of multiple-variable data sets.

Appendix A: Image processing system

The image processing system has two basic components, the host computer and the image processing workstation. The host performs data file management, selection of subimages, and initial data smoothing. The workstation provides interactive 3D imaging with rotation, data cutout, transparency, smoothing, and color selection. For this study the host computer is an IBM 3081. The workstation is an IBM 7350 Image Processing System. The IBM 7350 consists of a color display monitor, a conversational monitor, and a control unit. The control unit is channel-attached to the host under VM support. Images are formed in a refresh buffer; the buffer entries are transformed by a color lookup table and sent to the color monitor. The color monitor is a 1024 × 1024 CRT with 5 bits for each color (red, green, blue). The color lookup table has 4096 entries. (To change a colored range used for a particular structure, only 100 entries are changed, which permits rapid joystick interaction.) The 7350 used in this study has six working-storage buffers, each one megabyte; the refresh buffer is two megabytes. While transferring data between buffers, the Image Processor Arithmetic Logic Unit (IPALU) can be used to do 16-bit

arithmetic and Boolean operations on multiple buffer inputs. Buffer data can also be transformed with lookup tables before or after processing and transfer. The IBM 7350 has several window, mask, and selected-write features and a Random Address Processor that make it well suited to animation, 3D imaging, and multiple-window display. The display software MDIPS is written in VS FORTRAN.

● *Host functions*

As the frames are loaded into active memory, several operations can be performed. The first and last frame numbers to be loaded are specified, and a subimage can be selected. These two options can significantly reduce the storage used and the time required to form a 3D image in the 7350. The time required to form an image is directly proportional to the number of picture elements. The frames can be smoothed during loading. The electron and hole flux 2D magnitude arrays were smoothed with a 5 by 5 moving average. The resulting sequence of frames is saved in the 7350 storage buffers and is the base for subsequent imaging operations. The user interacts with the 3D data arrays in terms of physical dimensions, i.e., microns. At any time during the image display operations the user can request a new set of variables from the host. Also, the multiple-window image formed in the refresh buffer can be saved on the host disk, and previously stored buffers can be restored.

● *Workstation functions*

In the 7350 the 2D frames can be reviewed individually or in a rapid animated fashion. To form a 3D display on the color monitor, the user specifies the rotation of the input 3D array (colored ranges and direction lines) relative to an image plane. The display is then formed by projecting successive 2D frames of the data onto the image plane and by storing the result in the refresh buffer. The Random Address Processor in the 7350 control unit is used to map the data values in the frames onto the appropriate locations in the refresh buffer to form the projection. Vector fields can be displayed as direction lines or lines and magnitudes without any increase in display time, since the lines are stored as part of the input data array in the storage buffers. Since all of the frames are resident in the 7350 storage buffers, 3D images can be formed in 5 to 15 seconds of real time. The options of cutout, transparency, and contour lines do not increase the display time, since they are based on masking functions which are performed in parallel with the imaging. Multiple-window imaging occurs as rapidly as the user enters the parameters.

Animation is a powerful tool, but its value cannot be appreciated with a series of printed images. The dynamic visual presentation is an effective method of understanding the trends and changes in data. Also, the direct selection of the color of a region based on the displayed image needs to

be observed and tried to be appreciated. The direct, dynamic interaction with the data is an important capability.

Appendix B: Color selection

With a color monitor, ranges of data values can be displayed using different colors. Further, a color display has greater information capacity than a monochromatic display; each display point has three attributes with a color display (hue, lightness, and saturation) and only one attribute with a monochromatic display (lightness). However, color must be selected carefully for a clear and accurate display of the data.

In a 3D display, value ranges are displayed with a sequence of colors, dark in back and light in front. The best color sequence for a clear display of a range depends on its shape and relation to other colors. The user selects colors in terms of perceptual parameters: hue, lightness, and saturation (HLS). Two sets of HLS are required for each colored range, a rear set and a front set. Linear interpolation is used to form the color sequence for successive slices of the 3D display. In the current system, the user selects an end-point HLS with a joystick. The colored range and the front or rear end-point are selected. Joystick deflections in the x direction change hue and deflections in the y direction change lightness. Saturation is entered numerically; it is not very influential in improving shape perception. The image colors change rapidly with a joystick deflection, since the color lookup table can be changed rapidly. Joystick entry is very effective because colors are selected visually based on structure shapes and spatial relationships of colors.

References

1. T. E. Graedel and R. McGill, "Graphical Presentation of Results from Scientific Computations," *Science* **215**, No. 4537, 1191-1198 (1982).
2. L. T. Cook, S. J. Dwyer III, S. Batnitzky, and K. R. Lee, "A Three-Dimensional Display System for Diagnostic Imaging Applications," *IEEE Comput. Graph.* **3**, No. 5, 13-19 (1983).
3. E. Artzy, G. Frieder, and G. T. Herman, "The Theory, Design, Implementation and Evaluation of a Three-Dimensional Surface Detection Algorithm," *Comput. Graph.* **14**, No. 3, 2-9 (1980).
4. Edward J. Farrell, "Color Display and Interactive Interpretation of Three-Dimensional Data," *IBM J. Res. Develop.* **27**, No. 4, 356-366 (1983).
5. E. J. Farrell, R. Zappulla, and W. C. Yang, "Color 3D Imaging of Normal and Pathological Intracranial Structures," *IEEE Comput. Graph. & Appl.* **4**, No. 9, 5-17 (1984).
6. P. R. Atherton, "A Method of Interactive Visualization of CAD Surface Models on a Color Display," *Comput. Graph.* **15**, No. 3, 279-287 (1981).
7. J. N. Curtis and D. H. Schwieder, "Data Analysis Through a Generalized Interactive Computer Animation Method (DATICAM)," *Proceedings, Fourth Annual Conference of the National Computer Graphics Association*, 1983, pp. 601-608.
8. P. E. Cottrell and E. M. Buturla, "Two-Dimensional Static and Transient Simulation of Mobile Carrier Transport in a Semiconductor," *Proceedings, First International Conference on the Numerical Analysis of Semiconductor Devices (NASECODE I)*, Boole Press Limited, Dublin, Ireland, 1979, pp. 31-64.
9. E. M. Buturla, P. E. Cottrell, B. M. Grossman, C. T. McMullen, and K. A. Salsburg, "Three-Dimensional Transient Finite Element Analysis of the Semiconductor Transport Equations," *Proceedings, Second International Conference on the Numerical*

Analysis of Semiconductor Devices and Integrated Circuits (NASECODE II), Boole Press Limited, Dublin, Ireland, 1981, pp. 160-165.

10. E. M. Buturla, P. E. Cottrell, B. M. Grossman, and K. A. Salsburg, "Finite-Element Analysis of Semiconductor Devices: The FIELDAY Program," *IBM J. Res. Develop.* **25**, No. 4, 218-231 (1981).
11. Steven E. Laux, "Calculation of Quasi-Static Device Behavior with Small Computational Burden," *Proceedings, Third International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits (NASECODE III)*, Boole Press Limited, Dublin, Ireland, 1983, pp. 161-166.
12. B. M. Grossman, E. M. Buturla, and P. E. Cottrell, "Finite-Element Solution of the Semiconductor Transport Equations," *Computing Methods in Applied Science and Engineering*, J. L. Lyons and R. Glowinski, Eds., North-Holland Publishing Co., Amsterdam, the Netherlands, 1983.
13. R. H. Dennard, F. H. Gaensslen, L. Kuhn, and H. N. Yu, "Design of Micron Switching Devices," *Digest of Papers, International Electron Devices Meeting*, December 1972, p. 168.
14. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits* **SC-9**, No. 5, 256-268 (1974).
15. P. Richman, *MOS Field-Effect Transistors and Integrated Circuits*, John Wiley & Sons, Inc., New York, 1973.
16. S. M. Sze, *Physics of Semiconductor Devices*, 2nd Ed., John Wiley & Sons, Inc., New York, 1981.
17. T. Toyabe, K. Yamaguchi, S. Asai, and M. S. Mock, "A Numerical Model of Avalanche Breakdown in MOSFET's," *IEEE Trans. Electron Devices* **ED-25**, No. 7, 825-832 (1978).
18. A. Schuetz, S. Selberherr, and H. W. Poetzel, "A Two-Dimensional Model of the Avalanche Effect in MOS Transistors," *Solid-State Electron.* **25**, No. 3, 177-183 (1982).
19. A. Schuetz, S. Selberherr, and H. W. Poetzel, "Analysis of Breakdown Phenomena in MOSFET's," *IEEE Trans. Computer-Aided Design of Integrated Circuits & Systems* **CAD-1**, No. 2, 77-85 (1982).
20. D. S. Watanabe and S. Slamet, "Numerical Simulation of Hot-Electron Phenomena," *IEEE Trans. Electron Devices* **ED-30**, No. 9, 1042-1049 (1983).
21. G. A. Sai-Halasz and D. D. Tang, "Soft Error Rates in Static Bipolar RAMs," *Digest of Papers, International Electron Devices Meeting*, December 1983, pp. 344-347.
22. S. J. Polak, W. H. A. Schilders, C. D. Heijer, A. J. H. Wachters, and H. M. Vaes, "Automatic Problemsize Reduction for On-State Semiconductor Problems," *IEEE Trans. Electron Devices* **ED-30**, No. 9, 1050-1056 (1983).

Received September 17, 1984

Edward M. Buturla IBM Information Systems and Technology Group, P.O. Box 390, Poughkeepsie, New York 12602. Dr. Buturla is a senior engineer in the Mechanical Design Automation (MDA) Program Office, which is engaged in planning and strategies for a Corporate mechanical design system. He is currently responsible for

solid modeling and finite-element analysis applications. Previously he was with the Advanced Mathematics and Engineering Analysis Group in Burlington, Vermont, where he developed algorithms for semiconductor device simulation, electric and magnetic field modeling, and numerous mechanical disciplines. He received a B.S. in 1967 from the University of Connecticut, Storrs, and M.S. and Ph.D. degrees in mechanical engineering in 1970 and 1976, respectively, from the University of Vermont, Burlington. From 1976 to 1979, he was a Clinical Assistant Professor in the Department of Orthopedics and an Adjunct Professor in the Department of Mechanical Engineering, both at the University of Vermont. In 1977, he was selected by the National Society of Professional Engineers as the Young Engineer of the Year for the State of Vermont. Dr. Buturla received an IBM Outstanding Innovation Award in 1979 for his work in semiconductor modeling. He is a member of the Association for Computing Machinery, the American Society of Mechanical Engineers, the American Society of Biomechanics, Sigma Xi, and the National Society of Professional Engineers, and a registered professional engineer in the State of Vermont.

Phillip L. Corson IBM General Technology Division, Burlington facility, P.O. Box A, Essex Junction, Vermont 05452. Mr. Corson is a circuit design engineer currently working on high-performance CMOS static RAMs. He joined IBM in 1980 at Burlington; since then he has worked in product assurance and bipolar RAM development. His work assignments have included development work on a bipolar dynamic RAM, assignment of cell types to be used in high-performance, high-density RAM designs, and evaluation of the alpha sensitivity of static bipolar cells. Mr. Corson received his A.S. in electrical engineering from Southern Maine Vocational Technical Institute and is currently attending the University of Vermont.

Edward J. Farrell IBM Research Division, P.O. Box 218, Yorktown Heights, New York 10598. Mr. Farrell joined the Research Division as a research staff member in 1968. His current research is on image reconstruction and multidimensional color presentation of images and data. Prior studies involved the application of simulation and statistical methods to evaluate and predict the likelihood of recovery of critically ill patients. From 1966 to 1968, he was principal mathematician at UNIVAC, St. Paul, Minnesota. Mr. Farrell's work involved processing techniques and simulation related to spacecraft stabilization. From 1964 to 1966, he was a senior research scientist in the Research Division of Control Data Corporation. His research was concerned with feasibility and accuracy of different scanning optical sensors for celestial navigation. From 1959 to 1964, he was a mathematician in the Mathematics and Logic Research Department of UNIVAC, St. Paul. His assignments involved development of signal processing techniques related to sonar target classification and computer system reliability. Mr. Farrell received a B.S. from the University of Minnesota with a major in physics, and subsequently completed his doctoral course work in mathematical statistics. In 1982, he received an IBM Outstanding Technical Achievement Award for his work related to patient recovery.

Steven E. Laux IBM Research Division, P.O. Box 218, Yorktown Heights, New York 10598. Dr. Laux received the B.S. in electrical engineering from Purdue University, West Lafayette, Indiana, in 1976 and the M.S.E. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 1977 and 1981. Since graduation, he has been employed at the IBM Thomas J. Watson Research Center in Yorktown Heights in the silicon devices and technology area. His interests include MOS device modeling, applied mathematics, and device physics and technology. Dr. Laux is a member of the Institute of Electrical and Electronics Engineers.